Exploration par apprentissage de discussions de personnes en détresse psychologique

Remy Kessler, Nicolas Béchet, Gudrun Ledegen, Frédéric Pugniere-Saavedra

Exploration par apprentissage de discussions de personnes en détresse psychologique

Rémy Kessler[1], Nicolas Béchet[1], Gudrun Ledegen[2], Frédéric Pugnière-Saavedra[3]

[1] UNIV. BRETAGNE-SUD, UMR CNRS 6074, IRISA,
56017 Vannes, France
remy.kessler@univ-ubs.fr, nicolas.bechet@irisa.fr
[2] UNIV. RENNES II, PREFICS, EA 4246,
Campus Villejean
35043 Rennes, France
gudrun.ledegen@univ-rennes2.fr
[3] UNIV. BRETAGNE-SUD, PREFICS, EA 4246,
56017 Vannes, France
frederic.pugniere-saavedra@univ-ubs.fr

Résumé : Afin de s'adapter au mieux à la société, une association a développé une application de webchat permettant à toute personne d'exprimer et de partager ses préoccupations et ses malaises. Plusieurs milliers de conversations anonymes ont ainsi été réunies et forment un corpus inédit de récits sur la détresse humaine, les violences sociales. Nous présentons une méthode d'analyse de corpus combinant apprentissage non supervisé et *word embedding* afin de faire émerger les thématiques de cette collection particulière. Nous comparons la qualité de cette approche avec un algorithme standard de la littérature sur un corpus étiqueté et obtenons des résultats d'excellente qualité. Nous présentons une interprétation des regroupements obtenue sur cette collection particulière.

Mots-clés : word2vec, apprentissage non supervisé, word embedding

1 Introduction

Depuis les années quatre-vingt-dix, la souffrance sociale est une thématique qui fait l'objet d'une grande attention de la part de l'action publique et associative. Parmi les conséquences, figure l'explosion des lieux d'écoute ou des dispositifs sociotechniques de communication dont les finalités consistent notamment à modérer les diverses formes de souffrance par la libération de la parole dans un but thérapeutique Fassin (2004, 2006). Dans le cadre du projet METICS [1], une association de prévention du suicide a développé une application de *webchat* afin de répondre à ce besoin (Huet, 2015). Le *webchat* est un espace qui permet à toute personne d'exprimer et de partager avec un écoutant bénévole ses préoccupations et ses malaises. La principale spécificité de ce dispositif est son caractère non public et anonyme. Protégés par un pseudonyme, les écrivants sont invités à confier auprès d'un bénévole les aspects problématiques de leur existence. Plusieurs milliers de conversations anonymes ont ainsi été réunies et forment un corpus inédit de récits sur la détresse humaine. La finalité du projet METICS est de visibiliser les formes de souffrance ordinaires habituellement retranchées des espaces communs d'apparition et de saisir tant ses modalités d'énonciation que sa prise en charge au moyen des technologies numériques.

Dans le cadre de cette étude, nous souhaitons faire émerger de manière automatique les motifs de venue sur le chat des différents participants. En effet, même si l'association nous a fourni les thématiques abordées par l'ensemble des conversations (le travail, la solitude, la violence, le racisme, les addictions, les problèmes familiaux ou sentimentaux, etc.), le motif original de la venue sur le chat n'a pas été conservé. Dans la section suivante, nous présentons un état de l'art des différentes méthodes d'apprentissages comparables à notre proposition.

1. https://www.mshb.fr/projets_mshb/metics/2286/

La section 3 présente quelques statistiques sur les données tandis que la méthodologie est détaillée en section 4. La section 5 présente le protocole expérimental, une évaluation de notre système ainsi qu'une interprétation des regroupements finaux sur la collection de récits sur la détresse humaine.

2 Travaux connexes à notre proposition

La particularité de l'approche présentée dans ce papier est, d'un point de vue de l'utilisateur, de n'avoir à fournir que le libellé des classes à prédire. Ainsi, elle ne nécessite pas d'avoir un jeu de données étiquetées afin de prédire les différentes classes, c'est pourquoi elle est plus proche d'une méthode à base d'apprentissage non supervisé (ou semi-supervisé) que d'une méthode supervisée.

La littérature propose un certain nombre d'approches à base d'apprentissage non supervisé (ou clustering). L'idée du clustering est de regrouper des données non étiquetées dans un certain nombre de clusters, tel que des exemples similaires soient regroupés ensembles et ceux différents soient séparés. Pour une approche de clustering, le nombre de classes et la distribution des instances entre les classes ne sont pas connus *a priori* et le but est de trouver des regroupements significatifs. Les approches de clustering peuvent être classées selon le type de données fourni en entrée de l'algorithme et selon les critères de regroupement définissant la similarité ou la distance entre les données. Fraley & Raftery (1998) ont suggéré de diviser les algorithmes de clustering en deux catégories : les algorithmes hiérarchiques et ceux à base de partitionnement. Han & Kamber (2001) ont proposé de les catégoriser en trois catégories principales supplémentaires : les méthodes basées sur la densité, sur la modélisation et à base de grille.

L'algorithme de partitionnement des k-means fait partie des algorithmes de clustering les plus populaires, car il fournit un bon compromis entre la qualité de la solution obtenue et sa complexité de calcul (Arthur & Vassilvitskii (2007)). Même si k-means a été proposé pour la première fois il y a plus de 50 ans (MacQueen (1967)), il reste l'un des algorithmes les plus utilisés pour le clustering. En pratique, les k-means visent à trouver k centroïdes, un pour chaque cluster, minimisant la somme des distances de chaque instance de données par rapport à son centroïde respectif. Nous pouvons citer d'autres algorithmes à base de partitionnement comme les k-medoids ou PAM (Partition Around Medoids) qui est une évolution des k-means (Kaufman & Rousseeuw (1987)). Les approches hiérarchiques produisent quant à eux des clusters en partitionnant récursivement les données de manière descendante ou ascendante. Par exemple, dans une classification ascendante hiérarchique ou CAH (Lance & Williams (1967)), chaque exemple issu du jeu de données représente initialement un cluster. Ensuite, les clusters sont fusionnés, selon une mesure de similarité, jusqu'à ce que la structure arborescente souhaitée soit obtenue. Le résultat de cette méthode de clustering est appelé un dendrogramme.

Parmi les autres méthodes de clustering, les méthodes basées sur la densité supposent que les données appartenant à chaque cluster soient tirées d'une distribution de probabilité spécifique Banfield & Raftery (1993). L'idée est de faire croître un cluster donné tant que la densité dans le voisinage du cluster dépasse un certain seuil prédéfini. Les méthodes de classification basées sur un modèle reposent sur la découverte de descripteurs (ou caractéristiques) pour représenter chaque cluster. Les méthodes les plus utilisées pour ce type de méthodes sont les arbres de décision et les réseaux de neurones. Le plus populaire (qui sont à base de réseaux de neurones) sont les cartes de Kohonen ou self-organizing map - SOM (Kohonen (1982)). Finalement, les méthodes à base de grille partitionnent l'espace en un nombre fini de cellules qui forment une structure de grille.

Les approches à base d'apprentissage semi-supervisé tel que l'algorithme de propagation de libellés (Raghavan *et al.* (2007)) se rapprochent de la méthode proposée dans ce papier en se sens qu'elles consistent à utiliser un jeu de données d'apprentissage constitué de peu de données étiquetées et d'un nombre plus important de données non étiquetées afin de construire un modèle. Plus proche de la thématique de notre collection, Pestian *et al.* (2012) et Abboute *et al.* (2014) utilisent des approches supervisées pour détecter automatiquement

les personnes suicidaires dans les réseaux sociaux. Ils extraient des caractéristiques spécifiques pour entrainer différents classifieurs et compare les performances de leur système aux jugements de professionnels de la santé mentale. Plus récemment, une des tâches du challenge CLEF 2018 [2] était la détection des risques de dépression sur des textes écrits dans les médias sociaux Losada & Crestani (2016). Cependant, ces travaux et ce challenge impliquent des ensembles de données étiquetés, ce qui est la principale différence avec notre approche proposée (nous n'avons pas de jeu de données étiqueté).

3 Données et statistiques

L'association a fourni à l'équipe de recherche une collection de conversations entretenues entre les bénévoles et des appelants entre 2005 et 2015. La figure 1 présente un extrait anonymisé de conversation issue de cette collection.

```
...
Chat-association(21:35:55): Bonsoir
Appelant(21:34:14): Bsr, comment vivre avec un homme indecis?
Chat-association(20:32:33): c'est à dire?
Appelant(21:33:58): un homme qui veut divorcer un jour et un autre jour
qui est heureux d etre avec vous
Chat-association(20:34:35): Alors la question est peut-être que voulez-vous vous?
Appelant(21:35:54): je veux vivre heureu avec mon mari on vient de se marier
Appelant(21:37:03): donc je patiente j prends sur moi
Chat-association(20:38:12): lui comment il réagit?
Appelant(21:40:51): ms moi je suis son bouquet misere
Chat-association(20:41:12): "le bouquet misere"?
Appelant(21:42:17): bouc missaire
Chat-association(20:43:39): ok
Chat-association(20:44:59): vous avez le sentiment d'être un bouc émissaire
Appelant(21:55:40): oui
...
```

FIGURE *1 – Extrait d'une discussion issue de la collection METICS.*

Afin de réduire le bruit dans la collection, nous avons filtré l'ensemble des discussions contenant moins de 15 échanges entre un appelant et une personne de l'association, ces échanges étant généralement peu représentatifs (problème de connexion, demande d'information, etc.). Nous observons des phénomènes linguistiques bien particuliers comme des émoticônes [3], des apocopes (par exemple « ado », « télé », « bi ») des acronymes, des fautes (orthographiques, typographiques, mots collés, et d'une très grande morphovariabilité et d'une créativité explosive (Kessler *et al.*, 2004)). Ces phénomènes doivent leur origine au mode de communication (direct ou semi-direct), à la rapidité de composition du message ou aux contraintes technologiques de saisie imposées par le matériel (terminal mobile, tablette, etc.).

En complément de cette collection, nous avons utilisé dans le cadre de ces travaux un sous-ensemble du corpus des textes du journal Le-Monde [4]. Ce sous-ensemble issu de la collection d'origine contient les articles filtrés en fonction de leurs étiquettes thématiques. Nous conservons ainsi les articles ayant pour thématique la télévision, la politique, l'art, la science ou encore l'économie. Le tableau 1 présente quelques statistiques descriptives de ces deux collections.

4 Méthodologie

4.1 Vue d'ensemble du système

La figure 2 présente une vue d'ensemble du système dont les étapes seront détaillées dans le reste de la section. Au cours d'une première étape (module ①), nous appliquons différents

2. http://early.irlab.org/

3. Symboles utilisés dans les messages pour exprimer les émotions, exemple le sourire :-) ou la tristesse :-(

4. http://www.islrn.org/resources/421-401-527-366-2/

Collection	METICS	Le-Monde
Nombre total de documents	17 594	205 661
avant prétraitements linguistiques		
Nombre total de mots	12 276 973	87 122 002
Nombre total de mots différents	158 361	419 579
Nombre moyen de mots par conversation/doc.	698	424
après prétraitements linguistiques		
Nombre total de mots	4 529 793	41 425 938
Nombre total de mots différents	120 684	419 006
Nombre moyen de mots par conversation/doc.	257	201

TABLE *1 – statistiques du corpus METICS et du corpus Le-Monde.*

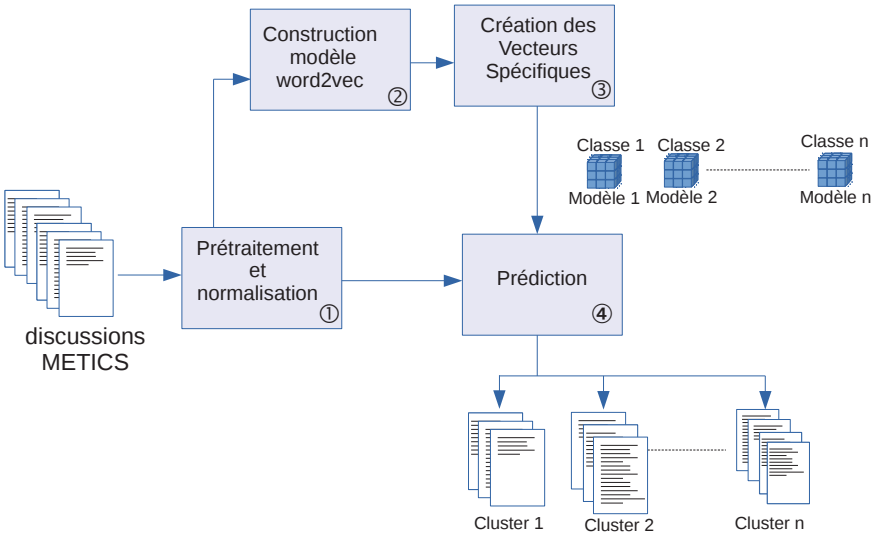

FIGURE *2 – Vue d'ensemble du système*

pré-traitements linguistiques à chacune des discussions. Le module suivant (②) constitue un modèle de *word embedding* à partir de ces discussions tandis que le troisième module (③) utilise ce modèle afin de créer des vecteurs spécifiques. Le dernier module ④ effectue une prédiction pour chaque discussion avant de séparer l'ensemble de la collection en regroupement en fonction de la classe prédite.

4.2 Pré-traitements et normalisation

Nous effectuons au préalable une extraction du contenu textuel de chaque discussion. Au cours de l'étape ①, une normalisation des textes est effectuée afin d'améliorer la qualité du processus. On supprime ainsi les accents, les caractères particuliers tels que « - », « / », « ( ) » . Différents processus linguistiques sont utilisés afin de réduire le bruit dans le modèle : les chiffres et nombres (numériques et/ou textuels), les symboles spéciaux ainsi que les termes contenus dans un anti-dictionnaire classique. Un processus de lemmatisation avait été intégré lors des premières expériences, mais il s'est avéré peu performant compte tenu de variations typographiques décrites en section 3. Au cours d'expériences préliminaires, nous avons tenté un filtrage permettant de prendre en compte uniquement l'un des rôles dans la conversation (appelant ou personne de l'association). Nous attribuons les faibles performances obtenues au rôle complémentaire de chaque intervenant (explicitation du message suivi d'une refor-

mulation) pour guider le système.

4.3 Construction d'un modèle word2vec

L'étape suivante de notre système consiste à construire un modèle de *word embedding* à l'aide de word2vec (Mikolov *et al.* (2013)). Cela consiste à projeter chaque mot de notre corpus dans un modèle vectoriel afin d'obtenir une représentation sémantique de ces derniers. Ainsi, des mots apparaissant dans des contextes similaires posséderont une représentation vectorielle relativement proche. Outre l'information sémantique, un des avantages d'une telle modélisation est le fait de produire des représentations vectorielles de mots, en fonction du contexte dans lesquels ils sont rencontrés. Ainsi, certains mots proches d'un terme *t* dans un modèle appris à partir d'un corpus *c1* pourront être très différents de ceux issus d'un modèle appris à partir d'un corpus *c2*. Nous observons par exemple sur la table 2 que les dix premiers mots proches du terme "ado" varient en fonction du corpus utilisé. Cet exemple montre également que l'utilisation d'un modèle générique type Wikipedia n'est pas pertinent dans notre cas, le corpus de l'association étant bruité et contenant un certain nombre d'apocopes, abréviations ou acronymes tels qu'"ado", "prob", ou encore "tv". Différents paramètres ont été testés et la configuration obtenant les meilleurs résultats a été conservée [5].

corpus	mots
METICS	adolescente, jeune, adolescence, 15ans, lycéenne, gaté, adolescent, gamin, gâtée, 14ans
Le-Monde	kyrou, fun, fm, nrj, marmelade, ouie, tropic, skyrock, difool, garnement

TABLE *2 – Dix mots les plus proches du terme "ado" selon le type de corpus en apprentissage*

4.4 Constitution des vecteurs spécifiques et prédiction

Au cours de cette étape, nous constituons des vecteurs contenant des termes choisis à l'aide du modèle word2vec construit au cours de l'étape 4.3. Pour chacune des thématiques contenues dans la collection, nous construisons ainsi un modèle linguistique spécifique en effectuant un plongement de mots (*word embedding*) afin de reconstruire le contexte linguistique de chacune des thématiques. Nous observons par exemple que les termes les plus proches de la thématique "travail" sont : « chômage », « boulot » , « boulo » , « stress » . De même, pour la thématique "addiction", nous observons les termes : « canabis », « alcoolisme » , « drogue » et « héroine » . Nous utilisons par la suite ce contexte afin de construire un vecteur, contenant la distance $dist_c(i)$ entre chaque terme i et la thématique c. Chacun de ces modèles étant indépendant, un même terme peut ainsi apparaître dans plusieurs modèles. On observe ainsi que le mot "stress" présent dans le vecteur "suicide" et dans celui de "travail", cependant le poids associé est différent. Nous avons fait varier la taille de ces vecteurs entre 20 et 1000 et les meilleurs résultats ont été obtenus avec une taille de 400. Au cours de la dernière étape ④, le système calcule un score S_c pour chaque discussion et pour chaque classe en fonction de chaque modèle linguistique tel que :

$$S_c(d) = \sum_{i=1}^{n} tf(i) \cdot dist_c(i) \tag{1}$$

avec i le terme considéré, $tf(i)$ la fréquence de i dans la collection, et $dist_c(i)$ est la distance entre le terme i et la thématique c. On attribue au final la classe ayant obtenu le score le plus élevé.

5. Les meilleurs résultats ont été obtenus avec les valeurs de paramètres suivantes : taille des vecteurs : 700, taille de la fenêtre glissante : 5, fréquence minimale : 10, méthode de vectorisation : skip-gram, utilisation d'une fonction softmax hiérarchique pour l'apprentissage du modèle.

5 Expériences et résultats

5.1 Protocole expérimental

Afin d'évaluer la qualité des clusters obtenus, nous avons utilisé un sous-ensemble des textes du journal Le-Monde, décrit en section 3, chaque article possédant une étiquette en fonction de la thématique. Dans le cadre de ces expériences, nous avons configuré l'approche des vecteurs spécifiques (VS) avec les paramètres optimaux, tels que définit en section 4.3 et 4.4. Afin de tester l'influence particulière de ce paramètre, nous avons également testé les vecteurs spécifiques sans la pondération. Afin de montrer la difficulté de la tâche, nous comparons notre système avec une *baseline* sous forme de tirage aléatoire, ainsi qu'avec l'algorithme des k-means (MacQueen (1967)), couramment utilisé dans la littérature tel que mentionné en section 2. Afin d'alimenter l'algorithme des k-means, nous avons transformé notre collection initiale en une matrice de type *bag of words* (Manning & Schütze (1999)) ou chaque conversation est décrire par la fréquence des mots qui la compose. Chacune des expériences a été évaluée en utilisant les mesures classiques de Précision, Rappel et F-score des documents bien classés, moyennés sur toutes les classes (avec $\beta = 1$ afin de ne privilégier ni la précision ni le rappel (Goutte & Gaussier (2005))). L'algorithme des k-means n'associant pas d'étiquette au regroupement produit, nous avons calculé de manière exhaustive l'ensemble des solutions pour ne garder que celle obtenant le F-score le plus élevé.

5.2 Résultats

	Precision	Rappel	F-score
Sans prétraitements linguistiques			
Baseline	0.18	0.16	0.17
k-means	0.23	0.20	0.22
VS sans pondération	0.54	0.50	0.52
Vecteurs Spécifiques (VS)	0.53	0.54	0.53
Avec prétraitements linguistiques			
k-means	0.30	0.21	0.25
VS sans pondération	0.55	0.51	0.53
Vecteurs Spécifiques(VS)	**0.54**	**0.54**	**0.54**

TABLE *3 – Ensemble des résultats obtenus par chaque système.*

Le tableau 3 présente une synthèse des résultats obtenus avec chaque système. On observe dans un premier temps que la baseline obtient un score très faible, mais qui reste relativement proche de l'aléatoire théorique (0,2) compte tenu du nombre de classes. L'utilisation des prétraitements linguistiques apporte peu individuellement, mais permet d'améliorer globalement les résultats des autres expériences. L'algorithme des k-means obtient des résultats légèrement meilleurs en termes de F-score, mais reste faible. Les vecteurs spécifiques obtiennent d'excellents résultats qui surpassent les autres systèmes avec un F-score de 0,54. L'exécution sans pondération montre que celle-ci permet d'améliorer légèrement le rappel.

5.3 Analyse des clusters

L'objectif initial de ses travaux étant l'exploration de la collection METICS, nous appliquons l'ensemble du processus avec l'approche des vecteurs spécifiques afin de catégoriser automatiquement l'ensemble des conversations du corpus. Dès lors, l'interprétation des clusters obtenus est réalisée avec l'Allocation de Dirichlet latente (ou latent Dirichlet allocation - LDA, Hoffman *et al.* (2010)) afin d'obtenir le sujet dominant de chaque cluster. Nous avons par la suite associé les poids à chacun des termes en fonction de chaque cluster et regroupé les mots-clés thématiques les plus significatifs dans le tableau 4. Ce dernier croise

les clusters avec deux types de matrices de discours : celles qui annoncent des directions sémantiques significatives en termes de présence (classées du rouge [très significatif] au bleu [moins significatif]) et celles qui reformulent globalement des éléments individuels. La méthode d'analyse utilisée semble opératoire puisque d'une part, sur 17 étiquettes de clusters préétablis, 10 comportent des désignations très pertinentes, 4 (*psy, adolescence, alcool*) le sont moins et seulement 3 (*handicap, travail, racisme*) ne le sont pas du tout. Le fait que ces trois dernières thématiques ne soient pas pertinentes, bien qu'ils soient des vecteurs de mal-être bien identifiés, est à mettre en lien avec le public particulier qui pratique le chat : pour environ 3/4 constitué de jeunes filles, le "travail" ne les concerne pas encore, et des thématiques "handicap" et "racisme" sont devancées par d'autres comme "solitude", "violence", "viol". Cette méthode d'analyse permet d'autre part de faire ressortir des univers sémantiques et des regroupements thématiques pour mettre en mots le mal-être et pour expliquer pourquoi il y a du mal-être chez le scripteur. L'entrée dans ce corpus par les clusters permet notamment de mettre au jour des embrayeurs fonctionnant comme des routines discursives significatives (Née *et al.*, 2014) qui annoncent et qui reformulent ce qui ne va pas.

	Comment dire "ce qui ne va pas"					
	du côté de l'annonce			du côté d'une forme de reformulation abstraite		
cluster	peur	psy	confiance	chose	difficile	problème
maladie	1,78	1,71	1,56		1,54	
adolescence	1,71	1,57	1,61	1,46	1,47	
solitude	1,69	1,64	1,58	1,52	1,55	0,22
suicide	1,67	1,71	1,54	1,51		
rupture	1,66	1,56	1,55	1,5	1,52	
violence	1,62	1,57	1,49	0,41	1,43	
travail	1,61	1,63	1,57	1,47	1,46	
viol	1,59	1,7	1,44	1,42	1,4	
angoisse	1,56	1,5	1,43	1,35	1,36	
famille	1,54	1,5	1,47	0,39	1	
relation	1,08	1	1,01	0,94	0,91	
alcool	0,89	0,88	0,27	0,79		0,5
deuil	0,88	0,96		0,77	0,77	
racisme	0,66	0,5		0,63		

TABLE *4 – Répartition des routines discursives par cluster.*

Dans la table 4, la peur, le psy et la confiance sont des désignations présentes pour chaque cluster avec un rang largement significatif ; pour autant le scripteur exprime-t-il toujours la peur quand il écrit, « j'ai peur d'être malade » ? Ces désignations ne participent-elles pas à ouvrir et à construire des sphères de significations autour de ces mots pivots ? Inversement, avec un rang inférieur, mais également significatif, les désignations chose, difficile, problème sont plus vagues, mais plus reformulantes pour reprendre les éléments qui participent à écrire ce qui ne va pas.

6 Conclusion et travaux futurs

Nous avons présenté dans cet article une approche non supervisée permettant d'explorer une collection de récits sur la détresse humaine. Cette approche utilise un modèle de *word embedding* afin de construire des vecteurs contenant uniquement du vocabulaire issu du contexte linguistique du modèle. Nous avons évalué la qualité de l'approche sur une collection étiquetée avec des mesures classiques. L'analyse détaillée a montré des résultats de très bonnes qualités (Fscore moyen de 0,54), comparativement aux autres systèmes testés. Cette méthode d'analyse a permis d'autre part de faire ressortir des univers sémantiques et des regroupements thématiques.

Nous envisageons dans un premier temps d'étudier plus en détail l'influence de chacun des paramètres sur les résultats obtenus. Nous envisageons par ailleurs afin de pouvoir attribuer plusieurs étiquettes à chaque discussion, ce qui permettrait de prendre en compte les chevauchements thématiques. L'analyse conforte l'approche par cluster pour faire ressortir les

traits définitoires de ce type de production de discours et pour en révéler un fonctionnement interne. Cette entrée par les routines discursives n'est qu'un exemple qui permettra ensuite d'aborder d'autres explorations avec notamment une focale sur les formes argumentatives et sur les formes d'intensité.

Références

ABBOUTE A., BOUDJERIOU Y., ENTRINGER G., AZÉ J., BRINGAY S. & PONCELET P. (2014). Mining twitter for suicide prevention. In *Natural Language Processing and Information Systems : 19th International Conference on Applications of Natural Language to Information Systems, NLDB 2014, Montpellier, France, June 18-20, 2014. Proceedings*, p. 250–253 : Springer International Publishing.

ARTHUR D. & VASSILVITSKII S. (2007). K-means++ : The advantages of careful seeding. *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, p. 1027–1035.

BANFIELD J. D. & RAFTERY A. E. (1993). Model-based gaussian and non-gaussian clustering. In *Biometrics*, volume 49, p. 803–821.

FASSIN D. (2004). Et la souffrance devint sociale. In *Critique, 680(1)*, p. 16–29.

FASSIN D. (2006). Souffrir par le social, gouverner par l'écoute. In *Politix, 73(1)*, p. 137–157.

FRALEY C. & RAFTERY A. E. (1998). How many clusters ? which clustering method ? answers via model-based cluster analysis. *The Computer Journal*, p. 578–588.

GOUTTE C. & GAUSSIER É. (2005). A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation. *ECIR 2005*, p. 345–359.

HAN J. & KAMBER M. (2001). Data mining : Concepts and techniques. *Kaufmann Publishers, USA*.

HOFFMAN M., BACH F. R. & BLEI D. M. (2010). Online learning for latent dirichlet allocation. In J. D. LAFFERTY, C. K. I. WILLIAMS, J. SHAWE-TAYLOR, R. S. ZEMEL & A. CULOTTA, Eds., *Advances in Neural Information Processing Systems 23*, p. 856–864.

HUET R. (2015). La voix des naufragés. Dire sa souffrance dans des associations d'écoute et de prévention du suicide. *Communication et langages. 2015/4 (186).*

KAUFMAN L. & ROUSSEEUW P. (1987). *Clustering by Means of Medoids*. Delft University of Technology : reports of the Faculty of Technical Mathematics and Informatics.

KESSLER R., TORRES J.-M. & EL-BÈZE M. (2004). Classification thématique de courriel par des méthodes hybrides. *Journée ATALA sur les nouvelles formes de communication écrite*.

KOHONEN T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, (1), 59–69.

LANCE G. N. & WILLIAMS W. T. (1967). A general theory of classificatory sorting strategies1. hierarchical systems. *The Computer Journal*, (4), 373–380.

LOSADA D. & CRESTANI F. (2016). A test collection for research on depression and language use. In *Proc. of Experimental IR Meets Multilinguality, Multimodality, and Interaction, 7th International Conference of the CLEF Association, CLEF 2016*, p. 28–39, Evora, Portugal.

MACQUEEN J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1 : Statistics*, p. 281–297, Berkeley, Calif. : University of California Press.

MANNING C. D. & SCHÜTZE H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA, USA : MIT Press.

MIKOLOV T., SUTSKEVER I., CHEN K., CORRADO G. & DEAN J. (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS'13*, p. 3111–3119, USA : Curran Associates Inc.

NÉE E., SITRI F. & VENIARD M. (2014). Pour une approche des routines discursives dans les écrits professionnels. *Congrès Mondial de Linguistique Française, DOI 10.1051/shsconf/20140801195.*

PESTIAN J. P., MATYKIEWICZ P., LINN-GUST M., SOUTH B., UZUNER O., WIEBE J., COHEN K. B., HURDLE J. & BREW C. (2012). Sentiment analysis of suicide notes : A shared task. *Biomedical Informatics Insights*, **5s1**, BII.S9042.

RAGHAVAN U. N., ALBERT R. & KUMARA S. (2007). Near linear time algorithm to detect community structures in large-scale networks. *Physical review. E, Statistical, nonlinear, and soft matter physics*, p. 036106.